Visualisierung hochdimensionaler Daten: State-of-the-Art von nichtlinearen Methoden zur Dimensionsreduktion

Tobias Schmieg, Helmut Beckmann[1]

Abstract: In vielen Forschungsdisziplinen ist die Visualisierung von hochdimensionalen Daten ein wichtiger Schritt der Datenanalyse. Eine Möglichkeit solche Daten zu visualisieren ist durch nichtlineare Methoden zur Dimensionsreduktion. In dieser Arbeit wird durch eine umfangreiche Literaturanalyse der State-of-the-Art von Methoden zur nichtlinearen Dimensionsreduktion ermittelt. Durch 51 untersuchten Publikationen werden 21 verschiedene Methoden identifiziert, wobei acht dieser Methoden näher betrachtet werden. Die älteren sechs dieser acht Methoden können zwar gut die lokale Struktur von hoher Dimensionalität in zwei oder drei Dimensionen darstellen, doch nur t-SNE (t-Distributed Stochastic Neighbor Embedding) und UMAP (Uniform Manifold Approximation and Projection for Dimension Reduction) können sowohl die globale als auch die lokale Struktur gleichzeitig beibehalten. Diese acht näher betrachteten Methoden stellen den aktuellen State-of-the-Art dar.

Keywords: nichtlineare Dimensionsreduktion, hochdimensionale Daten, State-of-the-Art, t-SNE, UMAP

1 Einleitung

Eine der vielfältigsten Techniken zur Visualisierung von Daten sind Scatterplots oder auch Streudiagramme genannt. Hierdurch können paarweise Korrelationen zwischen Dimensionen visualisiert werden. Diese Visualisierungstechnik kann, ebenfalls wie andere Techniken, höchstens dreidimensionale Daten darstellen. Durch eine paarweise Darstellung können zwar noch Korrelationen in Datensätzen mit mehr als drei Dimensionen untersucht werden. Jedoch werden dadurch auch nur die Zusammenhänge zwischen Dimensionen sichtbar, welche auch direkt in einer Visualisierung verglichen werden. Zusammenhänge über mehrere Dimensionen hinweg sind nur schwer zu erkennen [Ng20].

Jedoch ergeben sich immer mehr Anwendungsfälle, bei denen die Visualisierung hochdimensionaler Daten eine essenzielle Rolle spielt. Verschiedene Disziplinen, sowohl wissenschaftlich als auch in der Wirtschaft, nutzten Visualisierungen hochdimensionaler Daten,

1 Hochschule Heilbronn, Studiengang Wirtschaftsinformatik, Max-Planck-Str. 39, 74081 Heilbronn, {tschmieg@stud.; helmut.beckmann@}hs-heilbronn.de

https://doi.org/10.30844/AKWI_2022_13

um erste Erkenntnisse aus diesen zu ziehen. Diese sind unter anderem der medizinische Bereich, die Biowissenschaft oder auch die Analyse sozialer Netzwerke [Ha21]. So kommen Zellkerne, welche für die Brustkrebsforschung relevant sind, auf ca. 30 Variablen. Bei der Bildverarbeitung oder der Umgang mit Wort-Vektoren für Dokumentenverarbeitung kann die Anzahl der Dimensionen vierstellig oder mehr werden[vH08]. Daraus ergibt sich das Problem, dass nur eine gewisse Anzahl an unabhängigen Dimensionen gleichzeitig dargestellt werden kann [GS93].

Eine Lösung hierfür sind Methoden zur Dimensionsreduktion aus dem Fachgebiet des Machine Learning. Diese lassen sich in zwei Kategorien einordnen. Zum einen in lineare Methoden, welche versuchen, die Distanzen zwischen den Datenpunkten zu erhalten. Die wohl bekannteste Methode hierfür ist die Hauptkomponentenanalyse (Englisch: Principal Component Analysis (PCA)). Auf der anderen Seite gibt es noch die nichtlinearen Methoden. Diese versuchen durch nichtlineare Distanzen oder durch lokale Abstände die mehrdimensionalen Daten in einer niedrigeren Dimension darzustellen [Ho33]. Jedoch unterscheiden sich hier die Ansätze stark im Vorgehen und in dem genauen Ziel der Dimensionsreduktion. Die Anwendungsgebiete sind interdisziplinär und nicht auf Machine Learning beschränkt. Daraus ergibt sich folgende Forschungsfrage:

Welche Methoden sind State-of-the-Art für nichtlineare Dimensionsreduktion von hochdimensionalen Daten im Hinblick auf Datenvisualisierung?

Ziel dieser Forschungsarbeit soll es sein, den State-of-the-Art zu nichtlinearen Dimensionsreduktionsmethoden (NLDR-Methoden) darzustellen. Zur Beantwortung dieser Forschungsfrage werden in dieser Forschungsarbeit Publikationen untersucht, welche NLDR-Methoden anwenden, um das eigene Forschungsziel zu erreichen, oder Publikationen, welche NLDR-Methoden näher betrachten bzw. sogar eigene Ansätze vorstellen. Hierfür werden die im Rahmen einer umfangreichen Literaturrecherche identifizierten Publikationen auf die Nennung und Anwendung bzw. nähere Betrachtung von NLIDR-Methoden untersucht. Eine quantifizierte Darstellung der Untersuchungsergebnisse soll die Verbreitung von NLDR-Methoden sowie die Entwicklung des Forschungsgebietes über den betrachteten Zeitraum aufzeigen. Zur Erreichung dieser Zielsetzung werden nachfolgend die nötigen theoretischen Grundlagen gelegt. Aufbauend wird die Methodik sowie das Vorgehen zur Quantifizierung der Vorkommnisse genauer erläutert, welche im darauffolgenden Kapitel auf Basis des zuvor definierten Schemas analysiert wird, um Aussagen über die identifizierten NLDR-Methoden abzuleiten. Eine entsprechende Ergebnisdarstellung stellt die Erkenntnisse dieser Forschung am Ende dieser Forschungsarbeit zusammen. Hieraus resultierende Besonderheiten werden dediziert vorgestellt und bieten die Grundlage für weitergehende Forschung.

2 Dimensionsreduktion

Um einen Überblick über die theoretischen Grundlagen zu erhalten, wird der Begriff Dimensionsreduktion explizit dargestellt und erläutert. Die Nutzung von Dimensionsreduktion hat viele Anwendungsgebiete, darunter die Visualisierung hochdimensionaler Daten. Über die letzten Jahrzehnte kamen viele Techniken zur Visualisierung hochdimensionaler Daten auf. Doch viele dieser versuchten, lediglich mehr als zwei Dimensionen darzustellen. Die Interpretation wurde dem Nutzer der Visualisierung überlassen, was den eigentlichen Mehrwert in Frage stellt [vH08]. Deshalb besteht eine hohe Relevanz der Dimensionsreduktion für die Visualisierung hochdimensionaler Daten. Methoden hierfür lassen sich unter anderem in lineare und nichtlineare Methoden einteilen [LV07]. Bei PCA als vermutlich bekanntestes Beispiel der linearen Dimensionsreduktion, wird durch die Linearkombination von Variablen eines Datensatzes versucht möglichst viel Varianz in wenigen Dimensionen zu erhalten [Hu20] [Ha21]. Nichtlineare Methoden nutzen zu diesem Zweck nichtlineare Distanzmetriken oder die lokale Struktur zwischen Datenpunkten. So versucht Locally-Linear Embedding (LLE) [RS00] die Nachbardatenpunkte des hochdimensionalen Raumes ebenfalls als Nachbarn im niedrigdimensionalen Raum darzustellen. Andere Methoden wie Stochastic Neighbor Embedding (SNE) und darauf aufbauende Methoden, nutzen Wahrscheinlichkeiten, um die Abstände zwischen Datenpunkten darzustellen [Ha21]. Auch die Methoden t-SNE (t-Distributed Stochastic Neighbor Embedding) (welche eine der auf SNE aufbauende Methoden ist) und UMAP (Uniform Manifold Approximation and Projection for Dimension Reduction) nutzen für die Berechnung von Abständen Wahrscheinlichkeiten. Jedoch liegt bei diesen Methoden nicht nur die lokale Struktur von Datenpunkten im Fokus, sondern auch die globale Struktur des Datensatzes [vH08] [MHM18].

3 Methodik zur Literaturanalyse

Für das Vorgehen zur Beantwortung der Forschungsfrage wird eine umfangreiche Literaturanalyse gewählt. Nach Fettke [Fe06] wird eine fünfstufige Literaturanalyse durchgeführt, welche durch die Vorwärts- und Rückwärtssuche und das Erstellen einer Konzeptmatrix nach Webster und Watson [WW02] ergänzt wird.

Im Verlaufe dieser Forschungsarbeit werden die NLDR-Methoden herausgearbeitet. Dabei wird unterschieden, ob eine Methode nur genannt oder näher betrachtet wird. Betrachtet bedeutet dabei, dass die Funktionsweise oder Vor- und Nachteile der Methode beleuchtet werden oder die Methode angewendet wird. Außerdem wird bei der Untersuchung der Primärliteratur einer Methode, die Klassifizierung jener Methode in der Analyse nicht mit einbezogen, um eine dadurch entstehende Unschärfe zu vermeiden. Dies führt dazu, dass in der Konzeptmatrix (Abb. 1) pro Methode eine Klassifizierung Betrachtet häufiger vorzufinden ist, als in den folgenden Analysen. Die genaue Klassifizierung findet unter der Einschätzung der Autoren in Einbezug der oben genannten Punkte statt. Um die relevanten Methoden im Voraus zu filtern, werden Methoden, welche nur ein Vorkommnis aufweisen, nicht weiterverfolgt. Anschließend werden die Ergebnisse der Literaturanalyse in Form einer Konzeptmatrix untersucht, um die Forschungsfrage zu beantworten und den State-of-the-Art darzustellen.

4 Ergebnisse

Folgend werden die Ergebnisse der Literaturanalyse dargestellt. Mithilfe des Python-Visualisierungsframework matplotlib wird die Konzeptmatrix quantitativ aufbereitet.

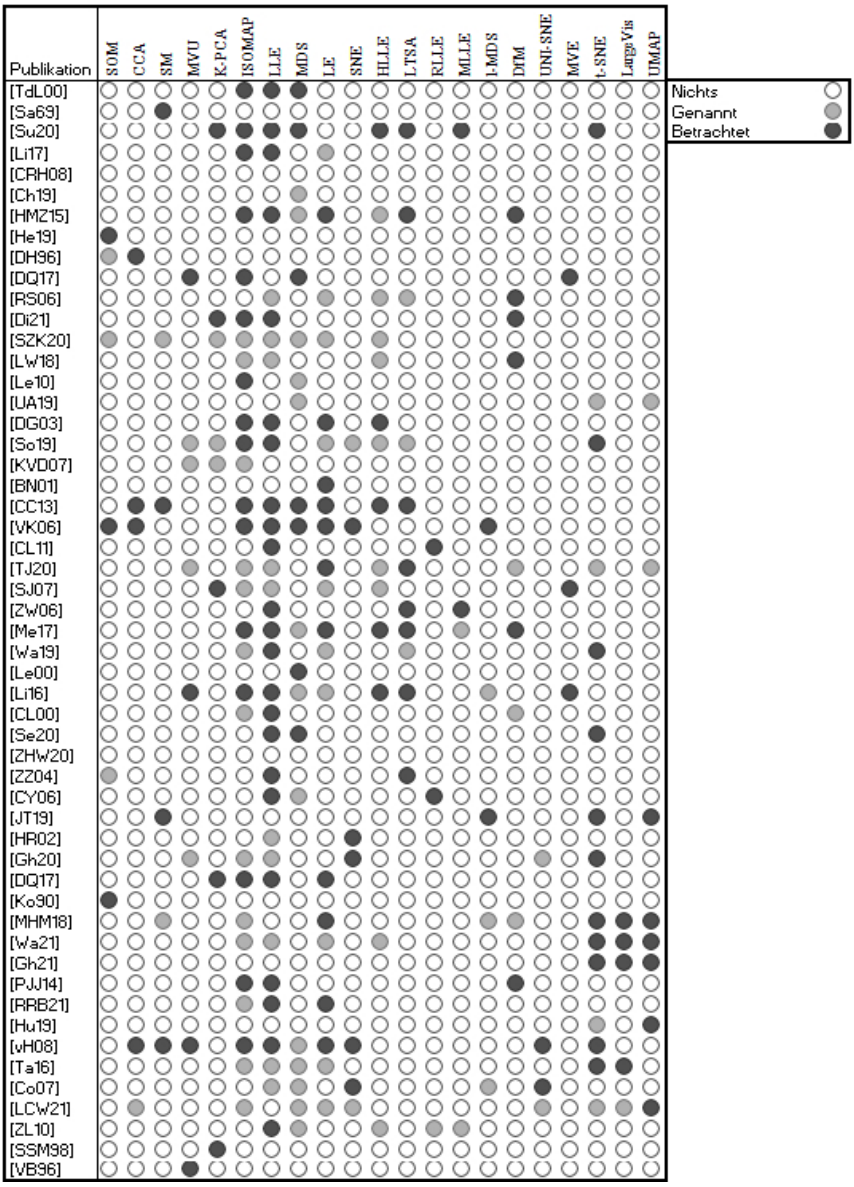

Abb. 1: Konzeptmatrix (eigene Darstellung)

Abb. 1 stellt die Konzeptmatrix dar, welche den Kern der Ergebnisse dieser Forschungsarbeit widerspiegelt. Zur Visualisierung der Vorkommnisse wird eine abgewandelte Form von Harvey-Balls verwendet. Die dreistufige Klassifizierung der Vorkommnisse innerhalb einer Publikation wird durch eine weiße Kugel für keine Nennung (Nichts), grau für Nennung (Nennung) und schwarz für Betrachtung (Betrachtet) wiedergegeben. Insgesamt werden 21 Methoden für die weitere Analyse miteinbezogen.

4.1 NLDR-Methoden

In Abbildung 2 werden die identifizierten NLDR-Methoden nach dem Veröffentlichungsjahr der dazugehörigen Primärquelle dargestellt. Im Durchschnitt hat eine NLDR-Methode ca. zehn Vorkommnisse. Auffälligkeiten zwischen den verschiedenen Methoden im Zusammenhang mit der Klassifizierung von genannt und betrachtet kann nicht festgestellt werden. Die Self-Organizing Map (SOM) [Ko90] ist mit dem Veröffentlichungsjahr 1990 die älteste und UMAP [MHM18] mit 2018 die jüngste NLDR-Methode. Trotz des Zeitraums von 28 Jahren machen die Veröffentlichungen im Zeitraum von 2000 bis 2004 (Isometric Mapping (ISOMAP) [TdL00], LLE [RS00], Mulitdimensional Scaling (MDS) [Le00], Laplacian Eigenmaps (LE) [BN01], SNE [HR02], Hessian LLE (HLLE) [DG03] und Local Tangent Space Alignment (LTSA) [ZZ04]) 63 % der Vorkommnisse aus.

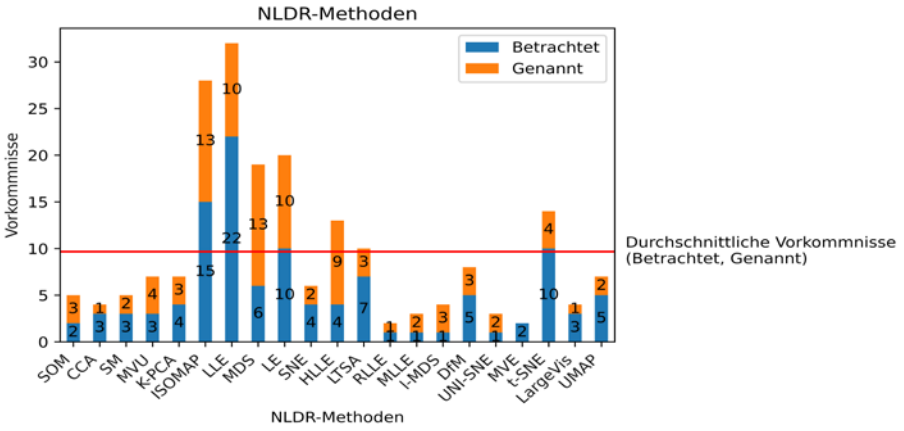

Abb. 2: Vorkommnisse der NLDR-Methoden (eigene Darstellung)

Zwar wurden einige Publikation der Datengrundlage in den Jahren direkt nach dem eben genannten Zeitraum veröffentlicht, trotzdem erschien ein Großteil ab 2017, wie in Abbildung 3 kenntlich wird.

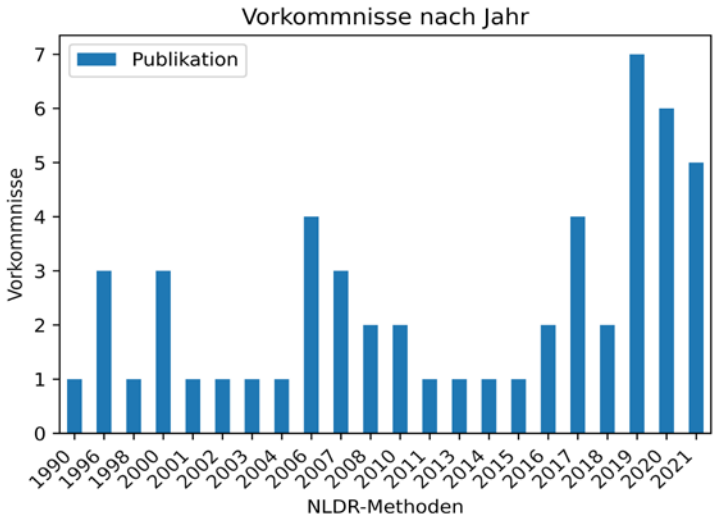

Abb. 3: Publikationen nach Veröffentlichungsjahr (eigene Darstellung)

Die nähere Betrachtung von NLDR-Methoden beschränkt sich deshalb im Folgenden auf ISOMAP, LLE, MDS, LE, HLLE, LTSA, t-SNE und UMAP. Bis auf UMAP weisen diese Methoden mindestens zehn Vorkommnisse auf. Da UMAP die jüngste Methode ist, kann für diese trotz lediglich sieben Vorkommnissen genügend Relevanz zur weiteren Betrachtung vorausgesetzt werden.

4.2 Nähere Betrachtung

Die acht NLDR-Methoden, welche für die weitere Betrachtung ausgewählt werden, machen 70 % der Vorkommnisse aus. Diese werden in Abbildung 4 dargestellt. In neueren Publikationen, wie die Primärliteratur zu t-SNE oder UMAP werden unter anderem ISOMAP und LE angewendet und entsprechend mit t-SNE und UMAP verglichen. Hier zeigt sich, dass ISOMAP, LE und weitere exemplarische Methoden nur schwer sowohl die lokale als auch die globale Struktur von Datensätzen gleichzeitig wiedergeben können. Weiterhin performen zwar alle angewendeten Methoden relativ gut auf künstlichen Datensätzen, jedoch bei realen Datensätzen wie MNIST zeigt sich die Schwäche der älteren Methoden [vH08], [MHM18]. Auch sind zu große Datensätze zum Teil ein limitierender Faktor für die älteren Methoden [MHM18]. Da sich LLE ebenfalls auf die direkten Nachbarn eines Datenpunktes fokussiert, hat diese Methode ähnliche Nachteile [RS00]. Zwar bietet HLLE Vorteile gegenüber LLE und anderen vorherigen NLDR-Methoden, trotzdem gilt es als robustere Variante der LEE Methode [DG03]. LTSA wendet Tangentenräume in der Nachbarschaft eines Datenpunktes an, um die lokale Geometrie darzustellen. Durch diese Tan-

gentenräume wird anschließend das globale Koordinatensystem konstruiert [ZZ04]. Doch auch hier liegt der Fokus auf den lokalen Nachbarschaften der Datenpunkte. Aussagen zu MDS sind schwer zu treffen, da das klassische MDS zwar eine lineare Methode ist, diese aber auch oft als Synonym für nichtlineare Abwandlungen wie beispielsweise local MDS genutzt wird.

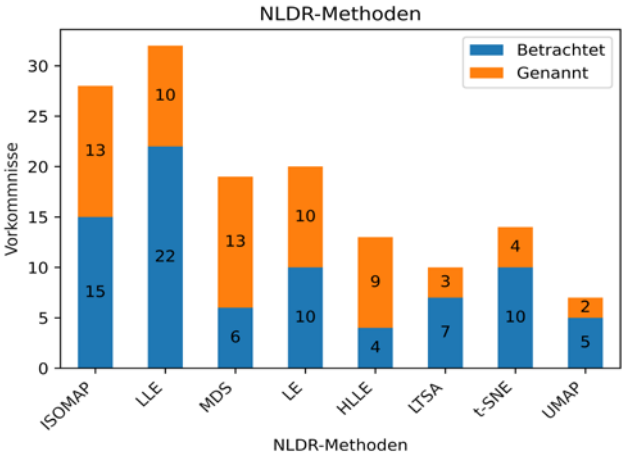

Abb. 4: Vorkommnisse der näher betrachteten NLDR-Methoden (eigene Darstellung)

In den Primärquellen zu t-SNE und UMAP wird bereits ein anderer Anspruch an die Visualisierungen, als die vorher diskutierten NLDR-Methoden deutlich. Während die globale Struktur der Datenpunkte in einer Visualisierung der Methoden ISOMAP, LLE, LE, HLLE und LTSA höchstens nebensächlich erwähnt wird, so wird dies bei t-SNE und UMAP neben der lokalen Struktur als klare Anforderung gestellt [vH08], [MHM18]. T-SNE basiert auf SNE, welches ebenfalls sechs Vorkommnisse aufzuweisen hat. Sowohl t-SNE als auch SNE geben die euklidischen Distanzen durch Wahrscheinlichkeiten wieder, um die Ähnlichkeiten zwischen Datenpunkten im niedrigdimensionalen Raum darzustellen. Jedoch nutzt t-SNE hierfür die studentische t-Verteilung, anstatt die Gauß-Verteilung. Außerdem wird für t-SNE eine effizientere Kostenfunktion angewendet, welche auch bei der Abwandlung Symmetric SNE genutzt wird. UMAP funktioniert im Kern relativ ähnlich zu t-SNE, jedoch konstruiert UMAP im hochdimensionalen Raum eine Repräsentation, welche genutzt wird, um die Repräsentation im niedrigdimensionalen Raum zu optimieren [MHM18].

4.3 Fazit

Es existiert eine Vielzahl an NLDR-Methoden. In dieser Arbeit wurden durch eine umfangreiche Literaturanalyse acht relevante Methoden identifiziert. Um Visualisierung von hochdimensionalen Daten zu erstellen, welche sowohl die lokale als auch die globale Struktur möglichst beibehalten, wird die Auswahl weiter eingeschränkt. Besteht nämlich solche Anforderung, so können von diesen acht Methoden lediglich t-SNE und UMAP diese im größeren Rahmen treffen. Zwar können t-SNE und vor allem UMAP nicht so viele Vorkommnisse in der Konzeptmatrix aufweisen, wie andere Methoden, jedoch kann hier das spätere Veröffentlichungsjahr als Hauptgrund angeführt werden.

5 Diskussion und Ausblick

Als Ergebnis dieser Forschungsarbeit können acht NLDR-Methoden genannt werden. Diese können die meisten Vorkommnisse im Untersuchungsgegenstand aufweisen. Jedoch ist hier anzumerken, dass die Vorkommnisse nicht direkt in Relevanz für den State-of-the-Art für NLDR-Methoden übertragen werden können. Weiterhin ist die Aussagekraft der Ergebnisse durch die Anzahl der untersuchten Publikationen von 51 Stück ebenfalls begrenzt. Zusätzlich haben ältere Methoden mehr Zeit gehabt Beachtung bei den Anwendern zu finden, während eventuelle besser performende Methoden nicht unbedingt schon die nötige Bekanntheit erreicht haben. Auf der anderen Seite konnte keine genauere Betrachtung der Methoden unter der Berücksichtigung verschiedener Anforderungen durchgeführt werden. So spielt es eine Rolle, ob eher die globale oder lokale Struktur für einen Anwendungsfall wichtig ist. Ebenso wurde die Performance im Hinblick auf Rechenzeit und -speicher nicht berücksichtigt. Auch die Datengröße und Anzahl von Dimensionen kann unterschiedliche Auswirkungen auf die Performance der verschiedenen Methoden haben. Zuletzt kommt noch hinzu, dass das Veröffentlichungsdatum der Primärquellen der Methoden nicht mit der Veröffentlichung der eigentlichen Methode übereinstimmen muss. Trotzdem können die Ergebnisse als Anhaltspunkt für den State-of-the-Art von NLDR-Methoden weiterverwendet werden.

Weiterführende Forschungsarbeiten könnten die acht NLDR-Methoden der Abb. 4 näher betrachten, um klare Aussagen über die Performance der Methoden treffen zu können. Außerdem könnte ein klares Kriteriumsraster oder Key-Performance-Indicator festgelegt werden, mit denen eine Evaluierung möglich ist. Weiterhin wäre interessant, inwieweit die älteren Methoden noch Relevanz aufweisen durch beispielsweise einfachere Implementierung oder Ähnliches.

Literaturverzeichnis

[BN01] Belkin, M.; Niyogi, P.: Laplacian eigenmaps and spectral techniques for embedding and clustering: Nips, S. 585–591, 2001.

[CC13] Chahooki, M. A. Z.; Charkari, N. M.: Learning the shape manifold to improve object recognition. Machine vision and applications 1/24, S. 33–46, 2013.

[Ch19] Chen, N. et al.: Application of computational intelligence technologies in emergency management: a literature review. Artificial Intelligence Review 3/52, S. 2131–2168, 2019.

[CL06] Coifman, R. R.; Lafon, S.: Diffusion maps. Applied and Computational Harmonic Analysis 1/21, S. 5–30, 2006.

[CL11] Chen, J.; Liu, Y.: Locally linear embedding: a survey. Artificial Intelligence Review 1/36, S. 29–48, 2011.

[Co07] Cook, J. et al.: Visualizing similarity data with a mixture of maps: Artificial Intelligence and Statistics, S. 67–74, 2007.

[CRH08] Carter, K. M.; Raich, R.; Hero III, A. O.: An Information Geometric Framework for Dimensionality Reduction, 2008.

[CY06] Chang, H.; Yeung, D.-Y.: Robust locally linear embedding. Pattern recognition 6/39, S. 1053–1065, 2006.

[DG03] Donoho, D. L.; Grimes, C.: Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. Proceedings of the National Academy of Sciences 10/100, S. 5591–5596, 2003.

[DH96] Demartines, P.; Herault, J.: CCA: “Curvilinear Component Analysis. 15° Colloque sur le traitement du signal et des images, 1995 ; p. 921-924, 1996.

[Di21] Ding, J.-E. et al.: Dopamine Transporter SPECT Image Classification for Neurodegenerative Parkinsonism via Diffusion Maps and Machine Learning Classifiers. arXiv preprint arXiv:2104.02066, 2021.

[DQ17] Ding, C.; Qi, H.-D.: Convex optimization learning of faithful Euclidean distance representations in nonlinear dimensionality reduction. Mathematical Programming 1/164, S. 341–381, 2017.

[Fe06] Fettke, P.: State-of-the-Art des State-of-the-Art. Wirtschaftsinformatik 4/48, S. 257, 2006.

[Gh20] Ghojogh, B. et al.: Stochastic neighbor embedding with Gaussian and Student-t distributions: Tutorial and survey. arXiv preprint arXiv:2009.10301, 2020.

[Gh21] Ghojogh, B. et al.: Uniform Manifold Approximation and Projection (UMAP) and its Variants: Tutorial and Survey. arXiv preprint arXiv:2109.02508, 2021.

[GS93] Gross, M. H.; Seibert, F.: Visualization of multidimensional image data sets using a neural network. The Visual Computer 3/10, S. 145–159, 1993.

[HA17] Henry, R. P.; Alfred, R.: Synergy in facial recognition extraction methods and recognition algorithms: International Conference on Computational Science and Technology, S. 358–369, 2017.

[Ha21] Haiyang Zhu et al.: Visualizing large-scale high-dimensional data via hierarchical embedding of KNN graphs. Visual Informatics 2/5, S. 51–59, 2021.

[He19] Hemmati, S. et al.: Bringing Manifold Learning and Dimensionality Reduction to SED Fitters. The Astrophysical Journal 1/881, L14, 2019.

[HMZ15] Hao, Z.-H.; Ma, S.-W.; Zhao, F.: Atlas compatibility transformation: A normal manifold learning algorithm. International Journal of Automation and Computing 4/12, S. 382–392, 2015.

[Ho33] Hotelling, H.: Analysis of a complex of statistical variables into principal components. Journal of educational psychology 6/24, S. 417, 1933.

[HR02] Hinton, G.; Roweis, S. T.: Stochastic neighbor embedding: NIPS, S. 833–840, 2002.

[Hu20] Hude, M. von der: Dimensionsreduktion - Hauptkomponentenanalyse englisch: principal components (PCA): Predictive Analytics und Data Mining Eine Einführung mit R. Springer Fachmedien Wiesbaden, Wiesbaden, S. 83–92, 2020.

[Hu19] Hurley, N. C. et al.: Visualization of Emergency Department Clinical Data for Interpretable Patient Phenotyping. arXiv preprint arXiv:1907.11039, 2019.

[JT19] Johannemann, J.; Tibshirani, R.: Spectral Overlap and a Comparison of Parameter-Free, Dimensionality Reduction Quality Metrics. arXiv preprint arXiv:1907.01974, 2019.

[Ko90] Kohonen, T.: The self-organizing map. Proceedings of the IEEE 9/78, S. 1464–1480, 1990.

[KVD07] Khurd, P.; Verma, R.; D., Christos (2007): Kernel-Based Manifold Learning for Statistical Analysis of Diffusion Tensor Images. In: (Karssemeijer, N.; Lelieveldt, B. Hrsg.): Information Processing in Medical Imaging. Berlin, Heidelberg, 2007. Berlin, Heidelberg: Springer Berlin Heidelberg, S. 581–593.

[LCW21] Liang, Y.; Chaudhuri, A.; Wang, H.: Visualizing the Finer Cluster Structure of Large-Scale and High-Dimensional Data: International Conference on Knowledge Science, Engineering and Management, S. 361–372, 2021.

[Le00] Leeuw, J. de: Multidimensional Scaling, 2000.

[Le10] Lei, Y. et al.: Fast ISOMAP Based on Minimum Set Coverage. In (Huang, De-Shuang and Zhang, Xiang and Reyes García, Carlos Alberto and Zhang, Lei Hrsg.): Advanced Intelligent Computing Theories and Applications. With Aspects of Artificial Intelligence. Springer Berlin Heidelberg, Berlin, Heidelberg, S. 173–179, 2010.

[Li16] Lin, T. et al.: Nonlinear Dimensionality Reduction by Local Orthogonality Preserving Alignment. Journal of Computer Science and Technology 3/31, S. 512–524, 2016.

[Li17] Lin Ma et al.: Adaptive Neighboring Selection Algorithm Based on Curvature Prediction in Manifold Learning, 2017.

[LV07] Lee, J. A.: Verleysen, M: Characteristics of an Analysis Method. In (Lee, J. A.; Verleysen, M. Hrsg.): Nonlinear Dimensionality Reduction. Springer New York, New York, NY, S. 17–45, 2007.

[LW18] Lin, C.-Y.; Wu, H.-T.: Embeddings of Riemannian manifolds with finite eigenvector fields of connection Laplacian. Calculus of Variations and Partial Differential Equations 5/57, S. 1–39, 2018.

[Me17] Mehta, K. et al.: Modified locally linear embedding with affine transformation. National Academy Science Letters 3/40, S. 189–196, 2017.

[MHM18] McInnes, L.; Healy, J.; Melville, J.: Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426, 2018.

[Ng20] Nguyen, Q. V. et al.: Evaluation on interactive visualization data with scatterplots. Visual Informatics 4/4, S. 1–10, 2020.

[PJJ14] Parekh, V. S.; Jacobs, J. R.; Jacobs, M. A.: Unsupervised nonlinear dimensionality reduction machine learning methods applied to multiparametric MRI in cerebral ischemia: preliminary results: Medical Imaging 2014: Image Processing, 90342O, 2014.

[RRB21] Ray, P.; Reddy, S. S.; Banerjee, T.: Various dimension reduction techniques for high dimensional data analysis: a review. Artificial Intelligence Review, S. 1–43, 2021.

[RS00] Roweis, S. T.; Saul, L. K.: Nonlinear dimensionality reduction by locally linear embedding. Science 5500/290, S. 2323–2326, 2000.

[Sa69] Sammon, J. W.: A Nonlinear Mapping for Data Structure Analysis. IEEE Transactions on Computers 5/C-18, S. 401–409, 1969.

[Se20] Seok, H.-S.: Performance comparison of dimensionality reduction methods on RNA-Seq data from the GTEx project. Genes & genomics 2/42, S. 225–234, 2020.

[SJ07] Shaw, B.; Jebara, T.: Minimum volume embedding: Artificial Intelligence and Statistics, S. 460–467, 2007.

[So19] Song, W. et al.: Improved t-SNE based manifold dimensional reduction for remote sensing data processing. Multimedia Tools and Applications 4/78, S. 4311–4326, 2019.

[Su20] Sun, Z. et al.: A Survey on Dimension Reduction Algorithms in Big Data Visualization, S. 375–395, 2020.

[SZK20] Shamai, G.; Zibulevsky, M.; Kimmel, R.: Efficient Inter-Geodesic Distance Computation and Fast Classical Scaling. IEEE Transactions on Pattern Analysis and Machine Intelligence 1/42, S. 74–85, 2020.

[Ta16] Tang, J. et al.: Visualizing Large-Scale and High-Dimensional Data: Proceedings of the 25th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, S. 287–297, 2016.

[TdL00] Tenenbaum, J. B.; de Silva, V.; Langford, J. C.: A Global Geometric Framework for Nonlinear Dimensionality Reduction. Science 5500/290, S. 2319–2323, 2000.

[TJ20] Ting, D.; Jordan, M. I.: Manifold Learning via Manifold Deflation. arXiv preprint arXiv:2007.03315, 2020.

[UA19] Urpa, L. M.; Anders, S.: Focused multidimensional scaling: interactive visualization for exploration of high-dimensional data. BMC bioinformatics 1/20, S. 1–8, 2019.

[vH08] van der Maaten, L.; Hinton, G.: Visualizing data using t-SNE. Journal of machine learning research 11/9, 2008.

[VK06] Venna, J; Kaski, S.: Local multidimensional scaling. Neural Networks 6/19, S. 889–899, 2006.

[Wa19] Wang, J. et al.: Multi-cancer samples clustering via graph regularized low-rank representation method under sparse and symmetric constraints. BMC bioinformatics 22/20, S. 1–15, 2019.

[Wa21] Wang, Y. et al.: Understanding how dimension reduction tools work: an empirical approach to deciphering t-SNE, UMAP, TriMAP, and PaCMAP for data visualization. J Mach. Learn. Res 22, S. 1–73, 2021.

[WW02] Webster, J.; Watson, R. T.: Analyzing the past to prepare for the future: Writing a literature review. MIS quarterly, S. xiii–xxiii, 2002.

[ZHW20] Zhang, S.; Huang, W.; Wang, Z.: Plant species identification based on modified local discriminant projection. Neural Computing and Applications 21/32, S. 16329–16336, 2020.

[ZL10] Zhang, S.-W.; Liu, J.: Weighted locally linear embedding for plant leaf visualization: International Conference on Intelligent Computing, S. 52–58, 2010.

[ZW06] Zhang, Z.; Wang, J.: MLLE: Modified Locally Linear Embedding Using Multiple Weights, S. 1593–1600, 2006.

[ZZ04] Zhang, Z.; Zha, H.: Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. SIAM journal on scientific computing 1/26, S. 313–338, 2004.